

Missing Data, Part II

Steyerberg, Chapter 7 and 9

Dave Glidden
27 October 2009

Last Week's Lab

- Fake Primary Biliary Cirrhosis Dataset
- Vars: Copper, Bilirubin, Spiders, Hepatom.
- 5,000 observations
- Spiders missing 50% of the time
- 3 scenario: also included actual spiders

True Model

$$\text{logit}(p_{\text{death}}) = \alpha + \beta_1 * \text{spiders} + \beta_2 * \text{hepatom} + \beta_3 * \text{copper_50} + \beta_4 * \text{bilir_2}$$

$$\exp(\beta_1) = \exp(\beta_2) = 1.5$$

$$\exp(\beta_3) = \exp(\beta_4) = 2.0$$

logistic regression

Complete Data

```
. logistic dead spiders hepat copp bil
```

```
Logistic regression
```

```
Number of obs = 5000
```

```
LR chi2(4) = 996.98
```

```
Prob > chi2 = 0.0000
```

```
Log likelihood = -2966.4366
```

```
Pseudo R2 = 0.1439
```

dead	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
spiders	1.979995	.1251107	10.81	0.000	1.749359	2.241038
hepatom	1.837629	.1157109	9.66	0.000	1.624276	2.079006
copper_50	1.502033	.0306591	19.93	0.000	1.443128	1.563341
bilir_2	1.523843	.0340713	18.84	0.000	1.458506	1.592106

Spiders MCAR

```
. logistic dead spider_mcar hepat copp bil
```

Logistic regression

```
Number of obs   =    2529  
LR chi2(4)      =    489.51  
Prob > chi2     =    0.0000  
Pseudo R2      =    0.1396
```

Log likelihood = -1508.0481

dead	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
spider_mcar	1.919294	.1704004	7.34	0.000	1.612759 2.284092
hepatom	1.768755	.1560701	6.46	0.000	1.487853 2.102691
copper_50	1.486381	.0419933	14.03	0.000	1.406313 1.571008
bilir_2	1.510922	.0464734	13.42	0.000	1.422527 1.604809

MCAR

- No bias
- Lost 50% of observations
- Relative efficiency of coef = 2.0
- Standard errors increased by $\sqrt{2}$ = 1.42
copper: $0.0420/0.0306 = 1.42$
- What if multiple missing variables?

Missingness Depends on Outcome

- Died: 75% spider data available
random subset
- Lived: 25% spider data available
random subset
- 50% overall missingness
- Available data mostly on those who die

Logistic Regression

```
. logistic dead spider_outcome hepat copp bil
```

```
Logistic regression
```

```
Number of obs = 2553
```

```
LR chi2(4) = 356.56
```

```
Prob > chi2 = 0.0000
```

```
Log likelihood = -1243.7473
```

```
Pseudo R2 = 0.1254
```

	dead	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
spider_out~e		1.76495	.1759302	5.70	0.000	1.451726	2.145756
hepatom		1.703542	.1698896	5.34	0.000	1.401086	2.071289
copper_50		1.464359	.0465622	12.00	0.000	1.375884	1.558523
bilir_2		1.508224	.052446	11.82	0.000	1.408856	1.6146

no bias. ever hear of a case control study?

Missingness on Predictors

- Spiders 70% measured based on
 - hepatom + & copper above median
 - hepatom - & copper below median
- Spiders 30% measured otherwise
- Spiders 50% overall missingness
- Missingness depends on predictors

Logistic Regression

```
. logistic dead spider_miss2 hepat copp bil
```

```
Logistic regression                               Number of obs   =    2474
                                                    LR chi2(4)      =    548.59
                                                    Prob > chi2     =    0.0000
Log likelihood = -1440.2882                       Pseudo R2      =    0.1600
```

```
-----+-----
      dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
spider_miss2 |   1.969892   .1786458     7.48  0.000     1.649106   2.353078
  hepatom    |   1.930353   .1799554     7.06  0.000     1.607993   2.317337
  copper_50  |   1.451152   .0434476    12.44  0.000     1.368446   1.538856
  bilir_2    |   1.559227   .0506123    13.68  0.000     1.463119   1.661649
-----+-----
```

no bias.

complete case unbiased if missing depends on preds

How about?

- If spiders is available on those who
 - died & had hepatomegaly
 - lived & had no hepatomegaly
- Missingness depends on predictors & outcomes
- This can't end well

Logistic Regression

```
. logistic dead spider_miss3 hepat bilir copper
```

```
Logistic regression                               Number of obs   =       2490
                                                    LR chi2(4)      =     1382.06
                                                    Prob > chi2     =       0.0000
Log likelihood = -1034.8263                       Pseudo R2      =       0.4004
```

```
-----+-----
      dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
spider_miss3 |   2.093397   .2352496     6.57   0.000     1.679562    2.609198
  hepatom   |  21.87057   2.545499    26.51   0.000    17.40962    27.47456
  bilir_2   |   1.561737   .061228    11.37   0.000     1.446227    1.686473
  copper_50 |   1.503422   .0543618    11.28   0.000     1.400563    1.613835
-----+-----
```

ummm....bias
but not in spiders -- in hepatomegaly. why?

Missingness

- Missingness associated with hep and death
- Took dead with hepat
- Alive without hepat
- Distorting association, creating bias
- But missingness is MAR:
depends on obs value death & hepat.

Complete Case

- Loss of efficiency:
 - MCAR
 - MAR: predictor only
- Bias:
 - MAR on predictor and/or outcome

We can do better

Multiple Imputation

1. Assume data MAR
2. Develop predictive model for missing data
3. Sample from predictive model
4. Make multiple complete datasets
5. Synthesize results

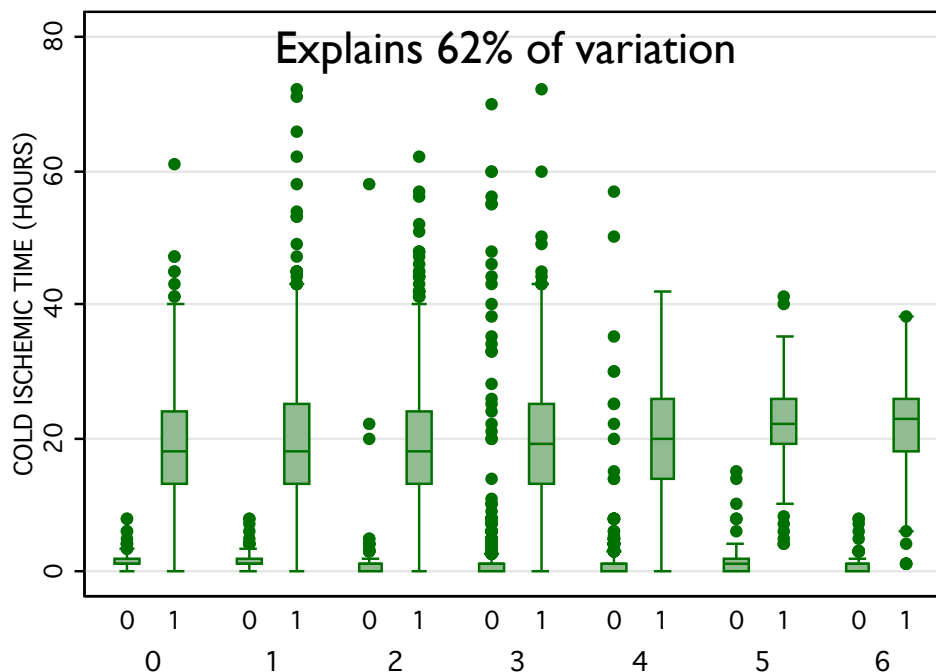
UNOS Data

Data Completeness

Variable	Obs	Mean	Std. Dev.	Min	Max
prevtx	9702	1.1443	.3514119	1	2
txtype	9775	.4733504	.4993148	0	1
age	9766	11.64653	5.291385	0	18
cold_isc	7525	10.85967	11.51735	0	72
hlatat	9541	2.57363	1.378919	0	6

prevtx: 1%
txtype: 0%
age: >1%
hla loci: 3%
cold_isc: 23%

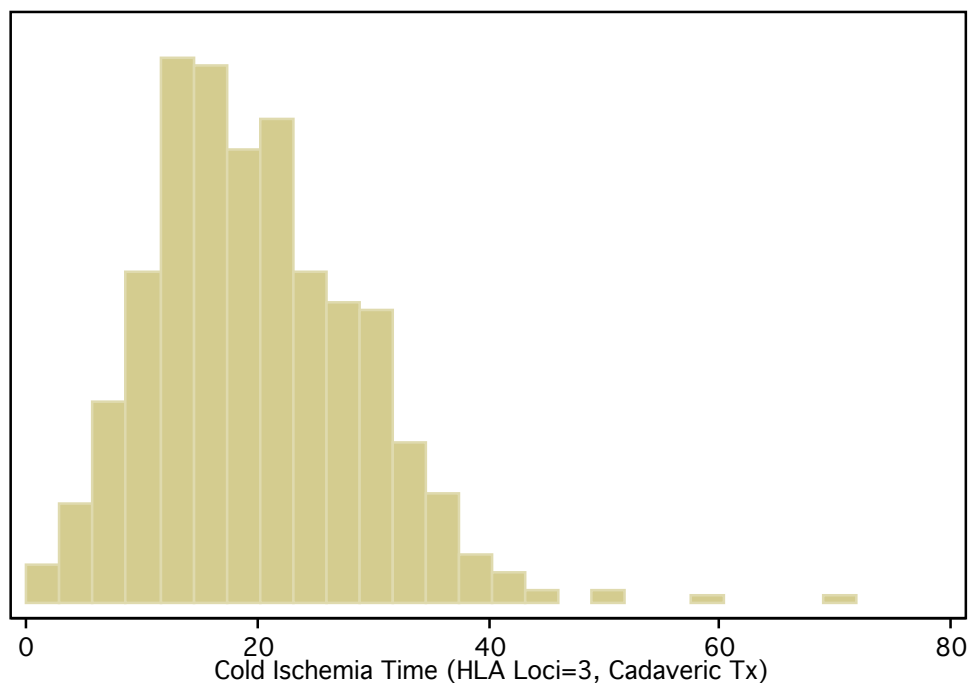
HLA and Donor Source



Two Steps

- Assume that cold_ischemia is MAR
- That values MAR conditional on HLA and txtype
- Txtype and # HLA loci can inform
- Predictive model might simply take cold_isc from dist of same loci and txtype

3 HLA, Cadaveric Tx



Good for motivating

Not the best in practice

- How confident in MAR: HLA, txttype if true, then there isn't any bias so why *not* include outcomes? and other things?
- Suddenly not easy to find matches
- Better distribution: base it on regression
- outcome: cold_isc, predictors: everything

Predictive Model

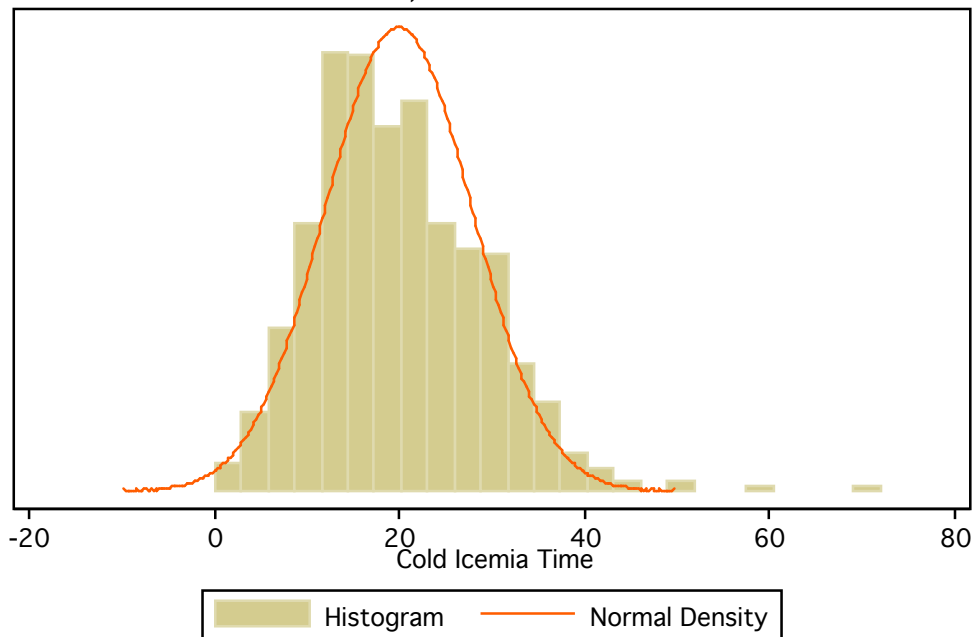
```
. xi: reg cold_isc txtty death i.hla age_don age year  
i.hlamat _Ihlamat_0-6 (naturally coded; _Ihlamat_0 omitted)
```

Source	SS	df	MS	Number of obs =	7347
Model	610095.267	11	55463.2061	F(11, 7335) =	1201.35
Residual	338636.724	7335	46.1672425	Prob > F	= 0.0000
Total	948731.991	7346	129.149468	R-squared	= 0.6431
				Adj R-squared	= 0.6425
				Root MSE	= 6.7946

cold_isc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
txttype	18.4926	.2161665	85.55	0.000	18.06885	18.91635
death	.3004369	.3743776	0.80	0.422	-.4334508	1.034325
_Ihlamat_1	.6360311	.3457854	1.84	0.066	-.0418076	1.31387
_Ihlamat_2	.4116543	.3433597	1.20	0.231	-.2614295	1.084738
_Ihlamat_3	.7623251	.3401847	2.24	0.025	.0954653	1.429185
_Ihlamat_4	.9979211	.3814195	2.62	0.009	.2502292	1.745613
_Ihlamat_5	1.650943	.5133476	3.22	0.001	.6446343	2.657252
_Ihlamat_6	1.748986	.5423876	3.22	0.001	.6857501	2.812221
age_don	.0170834	.0065371	2.61	0.009	.0042688	.0298979
age	.0295808	.0156351	1.89	0.059	-.0010686	.0602302
year	-.2436861	.0222149	-10.97	0.000	-.2872337	-.2001385
_cons	485.9413	44.36022	10.95	0.000	398.9825	572.9001

Predicted mean Cold isc = 19.9
SD about 6.8 (actually more)

HLA=3, Cadaveric Donor



Why Multiple Imputations?

- Don't know the value
- Single imputation: treats value as known
- Want to represent uncertainty
- Multiple draws from distribution

Imputation Model

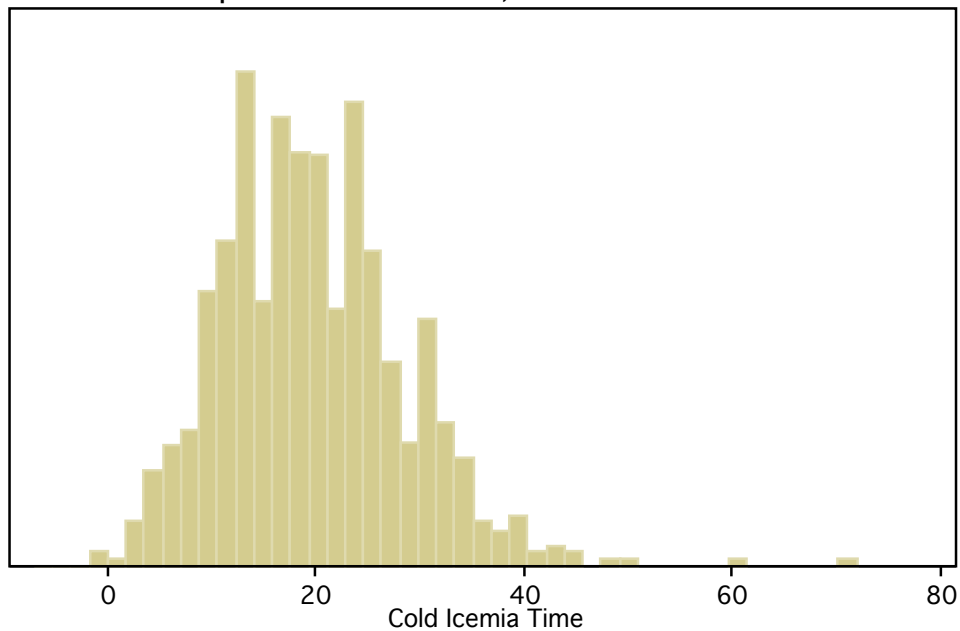
- Bias is the biggest threat
 - More predictors make MAR plausible
 - Include the outcome
- Unimportant covariates: little cost
 - probably not much gained
- Any predictor in model -- include impute
 - possible some that aren't

Source of Variation

- Predictive distn of cold_isc
 - $\text{cold_isc}_i = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_p * x_p + \epsilon_i$
- Sources of uncertainty
 1. prediction: ϵ_i
 2. parameters: $\beta_1, \beta_2, \dots, \beta_p$
 3. form of model: *predictors, transformation*
- Imputation deals with 1 & 2

Imputations

Imputations: HLA=3, Cadaveric Donor



Combining Imputations

$K = \text{Number of Imputations} = 5$

$\text{MI Estimate} = (OR_1 + OR_2 + \dots + OR_K) / K$

$\text{MI Variance} = (1 + 1/K) * \{ SE(OR_k) \}^2 +$
 $\{ \text{mean} (\{ SE_k(OR) \}^2) \}$

$\text{MI SE} = \text{sqrt}(\text{MI Variance})$

5 Imputations

```
. mi estimate, or : logistic death cold_isc txtx i.hlam age_don age year
```

```
Multiple-imputation estimates      Imputations =          5
Logistic regression              Number of obs =       9367
                                  Average RVI   =       0.0084
DF adjustment: Large sample      DF:      min   =       481.14
                                  avg         =       9.11e+08
                                  max         =       4.57e+09
Model F test:      Equal FMI      F( 11,446482.2)=      14.16
Within VCE type:   OIM           Prob > F       =       0.0000
```

death	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
cold_isc	1.000098	.0068917	0.01	0.989	.9866476	1.013732
txttype	1.292985	.2512163	1.32	0.186	.8833202	1.892645
hlamat						
1	1.003961	.1993305	0.02	0.984	.6803239	1.481555
2	.8206503	.1665123	-0.97	0.330	.5513748	1.221432
3	.6894886	.1429857	-1.79	0.073	.4592038	1.035258
4	.6344131	.1549459	-1.86	0.062	.3930782	1.023918
5	.8307535	.2621111	-0.59	0.557	.4476201	1.541824
6	.8066146	.2712919	-0.64	0.523	.4172325	1.559388
age_don	.9980962	.0039929	-0.48	0.634	.9903009	1.005953
age	.9589404	.0090785	-4.43	0.000	.9413111	.9769
year	.8557513	.0129072	-10.33	0.000	.8308239	.8814266

20 Imputations

```
. mi estimate, or : logistic death cold_isc txtx i.hlam age_don age year
```

```
Multiple-imputation estimates      Imputations =         20
Logistic regression              Number of obs =       9367
                                  Average RVI   =       0.0229
DF adjustment: Large sample      DF:      min   =       410.41
                                  avg         =       5.13e+08
                                  max         =       2.26e+09
Model F test:      Equal FMI      F( 11,340108.3)=      13.96
Within VCE type:   OIM           Prob > F       =       0.0000
```

death	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
cold_isc	1.001587	.007385	0.22	0.830	.9871749	1.01621
txttype	1.255866	.2527454	1.13	0.258	.8462613	1.863727
hlamat						
1	1.002362	.1990537	0.01	0.991	.6791866	1.479313
2	.819914	.166376	-0.98	0.328	.5508628	1.220375
3	.6884701	.1427788	-1.80	0.072	.4585199	1.033742
4	.6332496	.1546612	-1.87	0.061	.3923578	1.022039
5	.8281486	.2613009	-0.60	0.550	.4462041	1.537032
6	.8041512	.2704941	-0.65	0.517	.4159271	1.554742
age_don	.9980792	.0039913	-0.48	0.631	.9902869	1.005933
age	.9589001	.0090772	-4.43	0.000	.9412731	.9768572
year	.8561551	.0129298	-10.28	0.000	.8311844	.881876

50 Imputations

. mi estimate, or : logistic death cold_isc txyt i.hlam age_don age year

```

Multiple-imputation estimates          Imputations =          50
Logistic regression                   Number of obs =         9367
                                       Average RVI   =          0.0168
DF adjustment:  Large sample          DF:      min   =        1747.81
                                       avg         =        2.77e+09
                                       max         =        1.27e+10
Model F test:      Equal FMI          F( 11, 1.6e+06)=         14.06
Within VCE type:   OIM                Prob > F     =          0.0000
    
```

death	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
cold_isc	1.001891	.0071669	0.26	0.792	.9879328	1.016047
txtype	1.248634	.2474584	1.12	0.263	.8466672	1.841441
hlamat						
1	1.002063	.1989931	0.01	0.992	.678986	1.478869
2	.8197069	.1663335	-0.98	0.327	.5507243	1.220065
3	.6882046	.1427183	-1.80	0.072	.4583501	1.033327
4	.6330475	.1546022	-1.87	0.061	.3922444	1.021682
5	.8277018	.2611427	-0.60	0.549	.4459816	1.53614
6	.8035134	.2702734	-0.65	0.515	.4156034	1.553485
age_don	.9980769	.0039909	-0.48	0.630	.9902855	1.00593
age	.9588937	.0090768	-4.43	0.000	.9412675	.9768499
year	.8562408	.0129242	-10.28	0.000	.8312808	.8819502

500 Imputations

```

Multiple-imputation estimates          Imputations =          500
Logistic regression                   Number of obs =         9367
                                       Average RVI   =          0.0134
DF adjustment:  Large sample          DF:      min   =        26049.38
                                       avg         =        4.21e+10
                                       max         =        2.04e+11
Model F test:      Equal FMI          F( 11, 2.6e+07)=         14.11
Within VCE type:   OIM                Prob > F     =          0.0000
    
```

death	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
cold_isc	1.002185	.0070427	0.31	0.756	.9884757	1.016085
txtype	1.241517	.2439689	1.10	0.271	.844661	1.824833
hlamat						
1	1.001757	.1989317	0.01	0.993	.6787793	1.478415
2	.8195346	.1662986	-0.98	0.327	.5506085	1.219809
3	.6879439	.1426605	-1.80	0.071	.4581814	1.032925
4	.6328291	.1545431	-1.87	0.061	.392116	1.021312
5	.8272273	.2609825	-0.60	0.548	.445737	1.535221
6	.8029869	.2700856	-0.65	0.514	.4153419	1.552427
age_don	.9980732	.0039905	-0.48	0.630	.9902826	1.005925
age	.9588868	.0090764	-4.44	0.000	.9412615	.9768422
year	.8563155	.0129188	-10.28	0.000	.8313659	.8820139

Complete Case

```
. xi: logistic death cold_isc txty i.hlam age_don age year
i.hlamat      _Ihlamat_0-6      (naturally coded; _Ihlamat_0 omitted)
```

```
Logistic regression                               Number of obs   =       7347
                                                    LR chi2(11)    =       109.60
                                                    Prob > chi2    =       0.0000
Log likelihood = -1355.1632                       Pseudo R2      =       0.0389
```

death	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
cold_isc	1.001719	.0070634	0.24	0.808	.9879705 1.015659
txttype	1.314028	.2714015	1.32	0.186	.8765892 1.96976
_Ihlamat_1	1.116361	.2530693	0.49	0.627	.715887 1.740863
_Ihlamat_2	.9678571	.2218426	-0.14	0.887	.6176 1.516754
_Ihlamat_3	.8118071	.1910142	-0.89	0.376	.5118814 1.287468
_Ihlamat_4	.8065915	.2185684	-0.79	0.428	.4742386 1.371862
_Ihlamat_5	.9428916	.3421477	-0.16	0.871	.4630051 1.920162
_Ihlamat_6	.8938567	.3318802	-0.30	0.762	.4317428 1.850592
age_don	.9976235	.0042892	-0.55	0.580	.9892522 1.006066
age	.9694489	.0101854	-2.95	0.003	.94969 .9896188
year	.8669288	.0149339	-8.29	0.000	.8381475 .8966985

Relative Efficiency

Variable	RE
cold_isc	1.00
txttype	1.24
age_don	1.16
age	1.26
year	1.37

Comparison

- Big gains unlikely for missing variables
- Instead seen in other variables
- Deleting them bias and loss of efficiency
- Bias: certain covs patterns left out
recent years, living tx, younger recipients
- Efficiency: just lose numbers
- Imputed value not big diff here: not pred.

Did it work?

Compared to complete/available case analysis?

- No way to really know
- Must believe the MAR assumption
- No difference? Could be small % missing values
- Imputations close to overall mean for variables

How Many Imputations

- Surprisingly few
- 50% missingness 5 draws is highly efficient
- Stata on UNOS data
 - 20 imputations in 6 sec.
 - 50 imputations in 16 sec.
 - 100 imputation in 31 sec.
 - 500 imputations in 354 sec.

Making Up Data?

- No.
- Making a model for missingness
- Making guess at missing data
- Fully representing uncertainty
 - this is the key to it's validity

Issues in Modeling

- Most important: be inclusive
- More predictor: MAR more plausible
- Will remove biases -- primary objective
- Include outcome: other CC is unbiased
- Functional form not critical

mi package

- Powerful package for missing data
- Extensive capabilities
lots of choices appear esoteric
- Can handle multiple missing predictors
- Syntax is a little complicated

mi set

- Must declare data to be “mi set”
- `mi set mlong`
- Stores the imputations in long format
- I'd use as default
don't see advantage to another way

mi register

- Must declare variable as needing or not needing imputation
- `mi register imputed cold_isc`
- `mi register regular death hlamat
age_don age year`
- Says `cold_isc` will be imputed and others not

ICE Output

#missing values	Freq.	Percent	Cum.
0	1,893,947	99.89	99.89
1	2,086	0.11	100.00
Total	1,896,033	100.00	

Variable	Command	Prediction equation
cold_isc	regress	txttype death _Ihlamat_1 _Ihlamat_2 _Ihlamat_3 _Ihlamat_4 _Ihlamat_5 _Ihlamat_6 age_don age year
txttype		[No missing data in estimation sample]
death		[No missing data in estimation sample]
_Ihlamat_1		[No missing data in estimation sample]
_Ihlamat_2		[No missing data in estimation sample]
_Ihlamat_3		[No missing data in estimation sample]
_Ihlamat_4		[No missing data in estimation sample]
_Ihlamat_5		[No missing data in estimation sample]
_Ihlamat_6		[No missing data in estimation sample]
age_don		[No missing data in estimation sample]
age		[No missing data in estimation sample]
year		[No missing data in estimation sample]

mi impute

- This command actually does the imputations
- Can impute single or multiple variables
- Multiple capabilities appear quite limited *either all continuous variables or specific missingness patterns I'd not recommend it.*

Single Variable

- `mi impute regress cold_isc = death tctype i.hlam age_don age year , add(20) rseed(661)`
- Use linear regression to impute `cold_isc`
- Predictors: `death`, `tctype`, etc
- `add(20)` = 20 imputations
- `rseed()` = sets a random seed
*means you can replicate imputations
very very useful*

Imputed Regression

- `mi estimate, or : logistic death cold_isc tcty i.hlam age_don age year`
- Fits a logistic regression with imputations based on the last commands
- After “:” usual commands
- Need “or” to get odds ratios

MI Dataset

```
tabulate _mi_m
```

<u>_mi_m</u>	Freq.	Percent	Cum.
0	9,367	18.82	18.82
1	2,020	4.06	22.88
2	2,020	4.06	26.94
3	2,020	4.06	31.00
4	2,020	4.06	35.06
5	2,020	4.06	39.12
6	2,020	4.06	43.18
7	2,020	4.06	47.23
8	2,020	4.06	51.29
9	2,020	4.06	55.35
10	2,020	4.06	59.41
11	2,020	4.06	63.47
12	2,020	4.06	67.53
13	2,020	4.06	71.59
14	2,020	4.06	75.65
15	2,020	4.06	79.71
16	2,020	4.06	83.76
17	2,020	4.06	87.82
18	2,020	4.06	91.88
19	2,020	4.06	95.94
20	2,020	4.06	100.00
Total	49,767	100.00	

Passive Option

- Defines imputations for derived variables
 - transformation
 - interactions
- Original variables imputed
 - derived variables are passively imputed

Missing Values

- Important to explore and understand
- Informs model
- MAR permits valid analysis
missing depend on values thru observed data
- Missing data methods needed for validity and/or efficiency

Multiple Imputation

- One method for handling missing data
- Can be various others: have a similar spirit
averaging over distn for missing values
- Imputation is flexible and can be used in many problems
- Conceptually straightforward
- Lots of issues in practice